‘Quantum linguistics’ and Searle’s Chinese room argument

J. Mark Bishop, Slawomir J. Nasuto and Bob Coecke

Abstract Viewed in the light of the remarkable performance of ‘Watson’ - IBMs proprietary artificial intelligence computer system capable of answering questions posed in natural language - on the US general knowledge quiz show ‘Jeopardy’, we review two experiments on formal systems - one in the domain of quantum physics, the other involving a pictographic languaging game - whereby behaviour seemingly characteristic of domain understanding is generated by the mere mechanical application of simple rules. By re-examining both experiments in the context of Searle’s Chinese Room Argument, we suggest their results merely endorse Searle’s core intuition: that ‘syntactical manipulation of symbols is not sufficient for semantics’. Although, pace Watson, some artificial intelligence practitioners have suggested that more complex, higher-level operations on formal symbols are required to instantiate understanding in computational systems, we show that even high-level calls to Google translate would not enable a computer qua ‘formal symbol processor’ to understand the language it processes. We thus conclude that even the most recent developments in ‘quantum linguistics’ will not enable computational systems to genuinely understand natural language.

1 Background

“AS YOU read this article, your brain not only takes in individual words, but also combines them to extract the meaning of each sentence. It is a feat any competent reader takes for granted, but it’s beyond even the most sophisticated of today’s com-

J. Mark Bishop
Goldsmiths, University of London, UK, e-mail: bish@gold.ac.uk

Slawomir J. Nasuto
University of Reading, Reading, UK e-mail: s.j.nasuto@reading.ac.uk

Bob Coecke
University of Oxford, Oxford, UK e-mail: bob.coecke@cs.ox.ac.uk

puter programs. Now their abilities may be about to leap ahead, thanks to a form of graphical mathematics borrowed from quantum mechanics." So starts an article from The New Scientist[1] highlighting the work of Oxford University Computing Laboratory in quantum linguistics; a new approach to the study of language developed and explored by Bob Coecke, Mehrnoosh Sadrzadeh, Ed Grefenstette and Stephen Pulman (drawing from earlier work by Samson Abramsky and Bob Coecke on quantum computing). The article describes how the quantum and linguistics research groups at the Oxford University Computing Laboratory, are enabling computers to 'better understand' language by the application of the quantum picturalism formalism to linguistics; encoding words and grammar in a set of rules drawn from the mathematics of category theory. In this paper we investigate if 'quantum linguistics' genuinely enables computers to fully understand text.

2 Quantum physics

One morning in July 2011, at a meeting to discuss 'Foundational questions in the mathematical sciences', held at the International Academy in Traunkirchen, Austria, Bob Coecke from the University of Oxford, Slawomir Nasuto from the University of Reading and Mark Bishop from Goldsmiths College gathered over coffee[1] and discussed why it had taken more than sixty years from the birth of quantum physics to discover quantum teleportation. Bob suggested that the underlying reason was because 'Von Neumann Hilbert-space quantum mechanics' does not easily allow appropriate conceptual questions to be asked.

Bob subsequently outlined a radically new diagrammatic language - which he calls 'Quantum Picturalism' (QP) - so simple that it could be taught in kindergarten, but which is rich and powerful enough to facilitate simple derivations of relatively complex results in quantum physics. To illustrate its simplicity and power Bob explained that he has conceived an experiment involving school children which he anticipated would show quantum picturalism to be a 'language' powerful enough to derive complex phenomena in, say, quantum teleportation, but simple enough such that even kindergarten children could successfully use it with little or no prior knowledge of physics. But would these school children *really* be doing quantum physics we pondered over our coffee?

[1] There has since developed a serious dispute between the three participants as to if the discussion reported herein took place over coffee or over beer; or both.

3 Quantum picturalism

Quantum picturalism[6] defines a system consisting of formal operations[2] on a set of input/output (I/O) boxes connected by wires, which together define a QP picture (see Fig. 1).

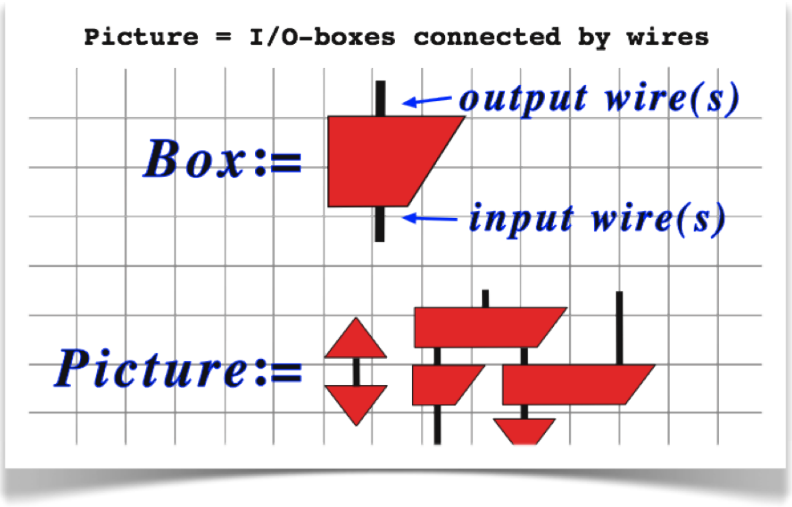

Fig. 1 *A 'Picture' in the quantum picturalism formalism*

Operations that can be performed on QP boxes include morphing and sliding: morphing entails transforming QP wires by stretching and constricting them; sliding boxes entails moving them around the image via 'sliding' them along the connecting wires (see Fig. 2). Substitution rules (see Fig. 3) define how one or more boxes can be replaced by another (or combined together or reduced/eliminated) to produce new picture elements.

Considering the QP diagram in Fig. 4, the box associated with the label 'Alice' can easily be moved (slid) across to align under the box associated with the label 'Bob'. Then, via the substitution rule shown in Fig. 3, both boxes can be combined and reduced to a basic wire. Thus, after the application of two simple rules we obtain a simplified QP diagram (on the right hand side of the equality) depicting Alice and Bob linked only by a wire.

[2] It could be argued, pace Wittgenstein on rule following[11], that such operations are not 'purely formal'; the boxes have 'meaningful tags' and require a primitive operational 'understanding' to follow the rules (e.g. to see that sliding to a specific position is 'OK'); however, as is shown in this work, on its own any minimal 'understanding' that accrues from formally manipulating QP elements in this way does not help ground the system in the target domain.

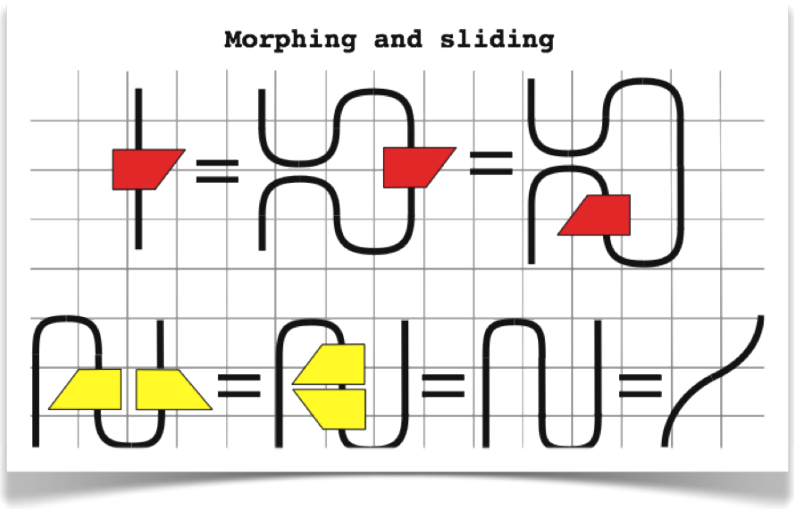

Fig. 2 *'Morphing' and 'Sliding' in the quantum picturalism formalism*

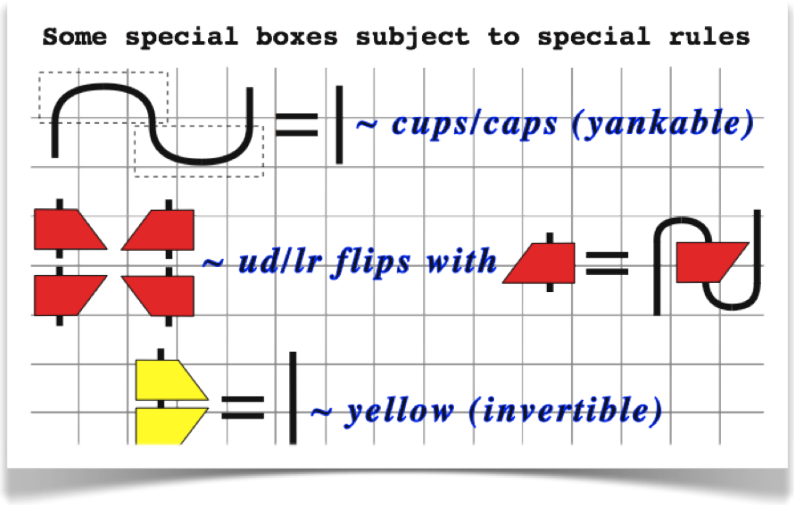

Fig. 3 *'Symbol substitution' in the quantum picturalism formalism*

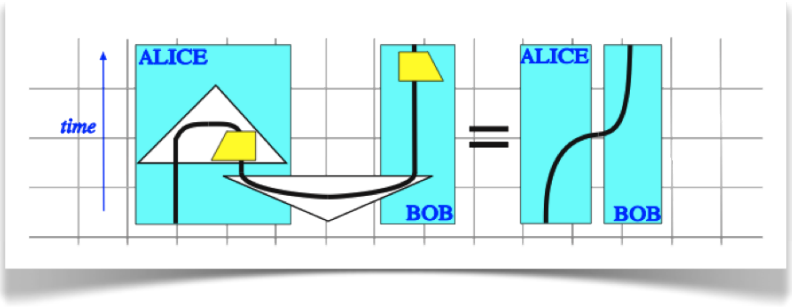

Fig. 4 *'Quantum teleportation' derived in the quantum picturalism formalism*

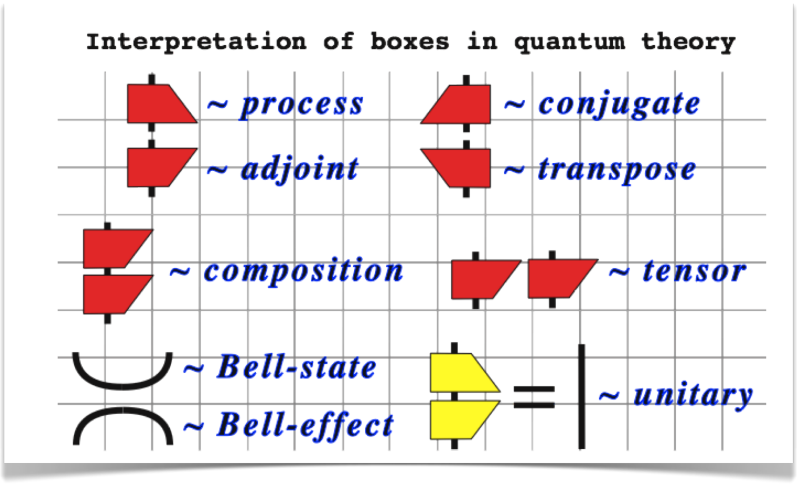

Fig. 5 *Symbol grounding in the quantum picturalism formalism*

Then, by applying the interpretation given in Fig. 5, we can understand the resulting QP diagram given in Fig. 4 in the context of the 'world of quantum physics' as meaning:

> "*Alice has an incoming quantum system (the input to the picture) and she and Bob also share a Bell-state (the white triangle with the cup inside. Alice then performs a certain operation on both of her quantum systems which depends on a unitary variable (the other white triangle where the box plays the role of the variable). Bob performs the conjugate to that unitary variable (the other box). The statement of equivalence with the right hand side then means 'At the end of performing the above stated instructions Alice's initially incoming quantum system will be with Bob'. This phenomenon is known as quantum teleportation*".

In demonstrating examples of QP in action (as above) Bob showed how even relatively simple formal operations on QP diagrams, in conjunction with understanding of the appropriate QP interpretation, can lead to new insights into the world

of quantum physics; insights (such as quantum teleportation) which may not be so obviously derived via classical application of Von Neumann Hilbert-space quantum mechanics. Reflecting again on Bob's proposed QP experiment with kindergarten children, we discussed just how deep an understanding, if any, of the QP *interpretation* is necessary for QP users to be *really* doing quantum physics? At this point in our discussion Slawomir recalled the work of Harré and Wang[8].

4 Is syntax sufficient for semantics?

In a brief paper from 1999 Harré and Wang described experiments with a simple pictorial 'language' comprising of a set of thirteen Chinese ideographs. Appropriate exchange of the symbols [between subjects competent in reading and writing Chinese ideographs] could facilitate very simple 'conversations' to take place: conversations[3] of the form:

Speaker-1 enquires: 'WHISKY??'
Speaker-2 replies: 'DRINK!'

Speaker-1 enquires: 'THIRSTY??'
Speaker-2 replies: 'BEER!'
Speaker-1 concludes: 'PUB..'

Harré and Wang subsequently developed and codified a simple set of 'purely formal' rules that could be used to automatically define appropriate responses for speaker-2 to make when passed symbols from speaker-1 (and vice versa). The rules of Harré and Wang's procedure described symbol transformations defined by a simple 'look-up table' (or 'rule-book') which encapsulated two types of response:

- '*Straight rules*' whereby, say, the symbol for 'WHISKY' is directly mapped to the symbol for 'DRINK'.
- '*Branching rules*' whereby, say, the symbol for 'THIRSTY?', if followed by the symbol for 'BEER!' maps to a response of 'PUB..'; but if followed by the symbol for 'COFFEE!' maps to a response of 'CAFE..'.

In their paper Harré and Wang's detail a series of experiments in which such 'iconic communication' was deployed between pairs of non-readers of Chinese, with the aim of determining if - by correctly iterating the application of the rule-book over time - non Chinese readers ever became able to ground [even primitive approximations to] the meanings of the Chinese ideographs. I.e. They evaluated precisely what a subject *actually experiences* in the context of a simple 'iconic languaging game' as a result of repeated low-level rule-based interactions.

[3] Readers in Ireland and the United Kingdom might recognise this style of conversation, so effectively deployed by Father Jack Hackett, in the Irish/British television comedy series *Father Ted.*

By stating in their conclusion that 'none of our participants reported having any sense of the meaning of the symbols', Harré and Wang's experiments demonstrated *in their experiments at least* that the iterated application of a small number of simple low-level rules to the manipulation of a small number of empty symbols, did not lead to the emergence of any understanding of what the symbols might refer to (mean); *that syntax is not sufficient for semantics*.

Of course the underlying claim - that syntax is not sufficient for semantics - is clearly conceptual and not empirical and hence its truth or falsity is not established by analysis of the Harré and Wang experiment described herein: as a reviewer of this paper trenchantly highlighted such a move would be analogous to claiming support for the conceptual philosophical assertion '*when a tree falls in the forest and no one is around to hear it doesn't make it sound*' by carrying out experiments on the particular cases of felling particular birch trees. However Mark recalled that the claim has been extensively conceptually probed by the American philosopher John Searle in his [now (in)famously] well known 'Chinese room' thought experiment, first published in the 1980 paper *Minds, Brains and Programs* (MBP)[9].

5 The Chinese room argument

Mark summarised Searle's Chinese Room Argument[4] (CRA) as follows[3]:

> "In 1977 Schank and Abelson published information[10] on a program they created, which could accept a simple story and then answer questions about it, using a large set of rules, heuristics and scripts. By script they referred to a detailed description of a stereotypical event unfolding through time. For example, a system dealing with restaurant stories would have a set of scripts about typical events that happen in a restaurant: entering the restaurant; choosing a table; ordering food; paying the bill, and so on. In the wake of this and similar work in computing labs around the world, some of the more excitable proponents of artificial intelligence began to claim that such programs actually understood the stories they were given, and hence offered insight into human comprehension.
>
> It was precisely an attempt to expose the flaws in the statements emerging from these proselytising AI-niks, and more generally to demonstrate the inadequacy of the Turing test[5], which led Searle to formulate the Chinese Room Argument.

[4] It is beyond the scope of this paper to summarise the extensive literature on the CRA other than to note that, to date, the two most widely discussed responses to the CRA have been the 'Systems reply' and the 'Robot reply'. For a broad selection of essays detailing these and other critical arguments see Preston and Bishop's edited collection '*Views into the Chinese room*'[2]. Conversely, by examining the application of the high-level quantum picturalism formalism to linguistics, this paper focuses on a response popular with some working within the fields of computing and artificial intelligence: that the 'purely formal' string-transformations defined in Searle's rule-book are both too simple and too low-level to ever facilitate the emergence of semantics and understanding.

[5] In what has become known as the 'standard interpretation' of the Turing test a human interrogator, interacting with two respondents via text alone, has to determine which of the responses is being generated by a suitably programmed computer and which is being generated by a human; if the interrogator cannot reliably do this then the computer is deemed to have 'passed' the Turing test.

The central claim of the CRA is that computations alone cannot in principle give rise to understanding, and that therefore computational theories of mind cannot fully explain human cognition. More formally, Searle stated that the CRA was an attempt to prove that syntax (rules for the correct formation of sentences:programs) is not sufficient for semantics (understanding). Combining this claim with those that programs are formal (syntactical), whereas minds have semantics, led Searle to conclude that 'programs are not minds'.

And yet it is clear that Searle believes that there is no barrier in principle to the notion that a machine can think and understand; indeed in MBP Searle explicitly states, in answer to the question 'Can a machine think?', that 'the answer is, obviously, yes. We are precisely such machines'. Clearly Searle did not intend the CRA to target machine intelligence *per se*, but rather any form of artificial intelligence according to which a machine could have genuine mental states (e.g. understanding Chinese) purely in virtue of executing an appropriate series of computations: what Searle termed 'Strong AI'.

Searle argues that understanding, of say a Chinese story, can never arise purely as a result of following the procedures prescribed by any computer program, for Searle offers a first-person tale outlining how he could instantiate such a program, and act as the Central Processing Unit of a computer, produce correct internal and external state transitions, pass a Turing test for understanding Chinese, and yet still not understand a word of Chinese.

Searle describes a situation whereby he is locked in a room and presented with a large batch of papers covered with Chinese writing that he does not understand. Indeed, the monoglot Searle does not even recognise the symbols as being Chinese, as distinct from say Japanese or simply meaningless patterns. Later Searle is given a second batch of Chinese symbols, together with a set of rules (in English) that describe an effective method (algorithm) for correlating the second batch with the first, purely by their form or shape. Finally he is given a third batch of Chinese symbols together with another set of rules (in English) to enable him to correlate the third batch with the first two, and these rules instruct him how to return certain sets of shapes (Chinese symbols) in response to certain symbols given in the third batch.

Unknown to Searle, the people outside the room call the first batch of Chinese symbols 'the script', the second set 'the story', the third 'questions about the story' and the symbols he returns they call 'answers to the questions about the story'. The set of rules he is obeying they call 'the program'. To complicate matters further, the people outside the room also give Searle stories in English and ask him questions about these stories in English, to which he can reply in English.

After a while Searle gets so good at following the instructions, and the 'outsiders' get so good at supplying the rules he has to follow, that the answers he gives to the questions in Chinese symbols become indistinguishable from those a true Chinese person might give.

From an external point of view, the answers to the two sets of questions, one in English the other in Chinese, are equally good; Searle, in the Chinese room, have passed the Turing test. Yet in the Chinese language case, Searle behaves 'like a computer' and does not understand either the questions he is given or the answers he returns, whereas in the English case, *ex hypothesi*, he does. Searle contrasts the claim posed by some members of the AI community - that any machine capable of following such instructions can genuinely understand the story, the questions and answers - with his own continuing inability to understand a word of Chinese; for Searle the Chinese symbols forever remain ungrounded[6]."

[6] The 'symbol-grounding' problem[7] is closely related to the problem of how words (symbols) get their meanings. On its own the meaning of a word on a page is 'ungrounded' and merely looking it up in a dictionary doesn't help ground it. If one attempts to look up the meaning of an unknown word in a [unilingual] dictionary of a language one does not already understand, one simply wanders endlessly from one meaningless definition to another (a problem not unfamiliar to young children); like Searle in his Chinese room, the search for meaning remains forever 'ungrounded'.

6 Complex rule-books

Historically, as Bob observed, Artificial Intelligence (AI) practitioners have been incredulous at the extreme simplicity of the low-level rules described by Searle (and deployed by Harré and Wang) that simply 'correlate one set of formal symbols with another set of formal symbols merely by their shape', such that typically very trivial combinations of un-interpreted symbols - Squiggles - map simply onto others - Squoggles. It has always seemed likely to such AI experts that any machine understanding program with a claim to real-world generality would require a very large and complex rule-base (program), typically applying very high-level rules (functions)[7].

However it is equally clear from MBP that Searle intended the CRA to be fully general - applicable to any conceivable [now or future] AI program (grammar based; rule based; neural network; Bayesian etc): '*I can have any formal program you like, but I still understand nothing*'. So if the CRA succeeds, it must succeed against even the most complex 'high-level' systems.

So, in a spirit of cooperation (between computer scientists, AI practitioners and Searle) let us consider a more complex formal program/rule-book-system which has (as one high-level-rule) a call to, say, Google-translate. We suggest that the internal representations scribbled on bits of paper used by the man in the room (monoglot Searle), could now maintain [at least partial] interpretations of the [unknown] Chinese text, as 'symbol-strings-in-English'.

In this way it is apparent that, via a process analogous to ones gradual understanding of a Chinese text via the repeated use of a Chinese-English dictionary, the application of [grounded] high-level-rules (Google-translate) to Chinese text would, over time, foster the emergence of genuine semantics and understanding in even a monoglot English speaker like Searle. Because both the rule-book and any internal representations the rule-book requires (Searle's 'scribbles on paper') are encoded in English, and *ex hypothesisi* Searle brings to the room an understanding of English, we suggest, pace Boden[4], that over time this *extended English Reply* would lead to the emergence of genuine semantics for Searle.

But does a computer Central Processing Unit[8] (CPU) really 'understand' its program and its variables [encoded as raw binary data] in a manner analogous to Searle's understanding of his rule-book and internal-representations encoded in English? In her 1988 paper (ibid) Maggie Boden suggests that, unlike say the human-driven manipulations of formal logic, it does; because, unlike the rules of logic, the execution of a computer program actually causes events to happen (e.g. it reads and writes data [or instructions] to memory and peripherals) and such 'causal seman-

[7] In contrast to the thirteen basic ideographs deployed by the Harré and Wang IBM's WATSON system - which recently won world wide acclaim as rivalling the greatest human players of the USA TV game show 'Jeopardy' - effectively deployed a complex high-level rule-book (literally thousands of complex algorithms working in parallel) on the full gamut of natural human language.

[8] A CPU is the core component of a computer system that executes program instructions (its algorithm or rule-book) by physically, and in most modern computers typically electronically, fetching or storing (reading or writing) them to and from memory and evaluating their coded commands.

tics' enable Boden to suggest that it is a mistake to regard [executing] computer programs as pure syntax and no semantics; such a CPU processing Chinese symbols really does have a 'toe-hold' on [Chinese] semantics. The analogy here is to Searle's understanding of the English language rule-book and hence the [extended, high-level] English reply holds.

In contrast to Boden we suggest, pace Wittgenstein[11], that the computer CPU does not really follow 'rules of its program' but merely acts in accordance to them; the CPU does not understand its internal-representations [as it executes its program and input] anymore than water in a stream 'understands' its flow down-hill; both are processes strictly entailed by their current state and that of the environment (their 'input').

Furthermore, pace Cassirer[5], we do not consider the computer as it executes its program with particular input(s) an 'information processor with a concomitant toe-hold in semantics, because we consider that the [physical] computer does not process symbols (which belong to the human realm of discourse), rather mere uninterpreted signals (binary digits [+/- 5v]) which belong to the world of physics.

'All syntax and no semantics' we suggest that, as there is no genuine sense in which the CPU understands its rule-book in a manner analogous to Searle's understanding of English, a CPU executing its program is simply not analogous to monoglot Searle's gradual understanding of a Chinese text via repeated use of an English/Chinese dictionary.

To reflect that the CPU merely mechanically transforms the signals it processes we simply insist, pace Searle, that the rule-book is defined only by syntactical operations (albeit perhaps more complex than the simple 'correlations' originally suggested by Searle and physically deployed by Harré and Wang) and the internal-representations ('scribbles on paper'), must remain defined by *un-interpreted* symbols (cf. Searle's 'Squiggles and Squoggles').

It is clear that, even allowing the rule-book to deploy high-level calls to, say Google-translate, because the internal-representations Searle is forced to manipulate remain mere un-interpreted signals (Squiggles and Squoggles), no understanding of the underlying Chinese text can ever emerge. The process is analogous to monoglot Searle's frustrated attempts to understand an unknown Chinese text using, say, only a Chinese/Japanese dictionary[9].

7 Quantum linguistics

This pioneering new approach to linguistics deploys quantum picturalism, the graphical form of *category theory*[10] originally developed for use in quantum mechanics and described earlier herein. Conventionally computers typically attempt

[9] Or Mark's lack of 'understanding' of quantum physics as he 'blindly follow the rules of QP with no concomitant understanding of an appropriate 'quantum physics' context; the QP interpretation.

[10] Category theory defines a branch of mathematics that allows different objects within a collection, or category, to be linked.

to 'understand' text as a collection of different words with limited structure; hence a computer may find it hard to tell the difference between 'Jane likes cheese' and 'Jane does not like cheese.' Conversely, despite the similarity of words in these sentences, their very distinct QP representations highlight their fundamental difference in meaning.

Bob likened the situation to watching a television program at the pixel level; 'rather than seeing the image, you get it in terms of 0s and 1s,' he says, and 'it wouldn't mean anything to you'. Similarly, by translating linguistic processes into the higher-level QP formalism, 'higher-level structures become visible'; in this manner quantum picturalism offers new insights, helping modern computational linguistic researchers develop ever more sophisticated natural language processing systems. Nonetheless, because at its heart the QP formalism merely offers computational linguistics a more complex (higher-level) rule-book, operating on more sophisticated - but still un-interpreted - QP representations, we suggest that any computational system qua 'quantum linguistics' remains as ignorant of the meaning of the text it processes as Searle is of Chinese.

8 Conclusion

At the end of our coffee-house journey from quantum picturalism to quantum linguistics via the Chinese room, we offer two modest observations made along the way:

- Unless they bring to Bob's proposed experiment relevant prior understanding of the QP interpretation in the world quantum physics (e.g. what a Bell-state is .., etc.), even if they discover a new result in quantum physics (e.g. quantum teleportation) kindergarten children cannot *really* be said to be doing quantum physics merely by correctly deploying the QP formalism.
- As syntax is not sufficient for semantics, even the mechanical execution of the high-level rule-book of quantum linguistics, deployed across the full gamut of natural language, will not result in a computational system genuinely capable of understanding the text it processes.

In Watson IBM finally put Searle's idealised component of the Chinese room (a complex program [rule-book] sophisticated enough to accurately respond to questions posed in natural language) to the test and in one sense (to the surprise of some) it passed; in Watson IBM have developed a system that [externally] exhibits astonishing [as-if] understanding/intelligence of the Jeopardy style questions it is posed. But would Searle, if he was ever locked in a 'Jeopardy room' and made to follow IBM's Watson rule-book, ever obtain understanding of playing the Jeopardy game? We conclude that - as syntax alone is never sufficient for semantics - he would not.

References

1. Aron J.: Quantum links let computers understand language. The New Scientist **208** 2790, 10–11 (2010)
2. Preston J. & Bishop J.M. (eds.) Views into the Chinese room. Oxford University Press, Oxford (2002)
3. Bishop J.M.: A view inside the Chinese room. Philosopher **28** (4), 47–51 (2004)
4. Boden M.: Escaping from the Chinese room. In: Boden M. (Ed.) The philosophy of Artificial Intelligence, pp. 89-105. Oxford University Press, Oxford (1988)
5. Cassirer E.: An Essay on Man. Yale University Press, New Haven (1944)
6. Coecke B.: Quantum Picturalism. Contemporary Physics **51**, 59–83 (2010)
7. Harnad S.: The Symbol Grounding Problem. Physica D **42**, 335–346 (1990)
8. Harré R. & Wang H.: Setting up a real 'Chinese room': an empirical replication of a famous thought experiment, Journal of Experimental & Theoretical Artificial Intelligence **11** (2), 153–154 (1999)
9. Searle J.R.: Minds, Brains, and Programs. Behavioral and Brain Sciences **3** (3), 417–457 (1980)
10. Schank R.C. & Abelson R.P.: Scripts, plans, goals and understanding: an inquiry into human knowledge structures. Erlbaum, Hilsdale NJ (1977)
11. Wittgenstein L.: Philosophical Investigations. Blackwell, Oxford UK (1958)